An Extended Overview of the CLEF 2020 ChEMU Lab: Information Extraction of Chemical Reactions from Patents

Jiayuan He[1], Dat Quoc Nguyen[1,8], Saber A. Akhondi[2], Christian Druckenbrodt[3], Camilo Thorne[3], Ralph Hoessel[3], Zubair Afzal[2], Zenan Zhai[1], Biaoyan Fang[1], Hiyori Yoshikawa[1,5], Ameer Albahem[4], Jingqi Wang[6], Yuankai Ren[7], Zhi Zhang[7], Yaoyun Zhang[6], Mai Hoang Dao[8], Pedro Ruas[9], Andre Lamurias[9], Francisco M Couto[9], Jenny Copara[10,11,12], Nona Naderi[10,11], Julien Knafou[10,11,12], Patrick Ruch[10,11], Douglas Teodoro[10,11], Daniel Lowe[13], John Mayfield[14], Abdullatif Köksal[15], Hilal Dönmez[15], Elif Özkırımlı[15,16], Arzucan Özgür[15], Darshini Mahendran[17], Gabrielle Gurdin[17], Nastassja Lewinski[17], Christina Tang[17], Bridget T. McInnes[17], Malarkodi C.S.[18], Pattabhi Rk Rao.[18], Sobha Lalitha Devi[18], Lawrence Cavedon[4], Trevor Cohn[1], Timothy Baldwin[1], and Karin Verspoor[1(✉)]

[1] The University of Melbourne, Melbourne, Australia
{estrid.he,hiyori.yoshikawa,trevor.cohn,tbaldwin,karin.verspoor}@unimelb.edu.au
{zenan.zhai,biaoyanf}@student.unimelb.edu.au
[2] Elsevier BV, Amsterdam, The Netherlands
[3] Elsevier Information Systems GmbH, Frankfurt, Germany
{s.akhondi,c.druckenbrodt,c.thorne.1,r.hoessel,m.afzal.1}@elsevier.com
[4] RMIT University, Melbourne, Australia
{ameer.albahem, lawrence.cavedon}@rmit.edu.au
[5] Fujitsu Laboratories Ltd., Japan
[6] Melax Technologies, Inc, Houston, USA
[7] Nantong University, Nantong, China
{jingqi.wang,yaoyun.zhang}@melaxtech.com
[8] VinAI Research, Vietnam
{v.datnq9, v.maidh3}@vinai.io
[9] LASIGE, Universidade de Lisboa, Lisbon, Portugal
{psruas, fcouto}@fc.ul.pt, alamurias@lasige.di.fc.ul.pt
[10] Uni. of Applied Sciences & Arts of Western Switzerland, Geneva, Switzerland
[11] Swiss Institute of Bioinformatics, Geneva, Switzerland
[12] University of Geneva, Geneva, Switzerland
{jenny.copara,nona.naderi,julien.knafou,patrick.ruch,douglas.teodoro}@hesge.ch
[13] Minesoft, Cambridge, United Kingdom
[14] NextMove Software, Cambridge, United Kingdom
daniel@minesoft.com, john@nextmovesoftware.com
[15] Boğaziçi University, Istanbul, Turkey
[16] Data and Analytics, F. Hoffmann-La Roche AG, Switzerland
{abdullatif.koksal,hilal.donmez,elif.ozkirimli,arzucan.ozgur}@boun.edu.tr
[17] Virginia Common Wealth University, Richmond, United States
{mahendrand,gurding,nalewinski,ctang2,btmcinnes}@vcu.edu
[18] MIT Campus of Anna University, Chennai, India
csmalarkodi@gmail.com, {pattabhi, sobha}@au-kbc.org

Abstract. The discovery of new chemical compounds is perceived as a key driver of the chemistry industry and many other economic sectors. The information about the new discoveries are usually disclosed in scientific literature and in particular, in chemical patents, since patents are often the first venues where the new chemical compounds are publicized. Despite the significance of the information provided in chemical patents, extracting the information from patents is costly due to the large volume of existing patents and its drastic expansion rate. The Cheminformatics Elsevier Melbourne University (ChEMU) evaluation lab 2020, part of the Conference and Labs of the Evaluation Forum 2020 (CLEF2020), provides a platform to advance the state-of-the-arts in automatic information extraction systems over chemical patents. In particular, we focus on extracting synthesis process of new chemical compounds from chemical patents. Using the ChEMU corpus of 1500 "snippets" (text segments) sampled from 170 patent documents and annotated by chemical experts, we defined two key information extraction tasks. Task 1 targets at chemical named entity recognition, i.e., the identification of chemical compounds and their specific roles in chemical reactions. Task 2 targets at event extraction, i.e., the identification of reaction steps, relating the chemical compounds involved in a chemical reaction. In this paper, we provide an overview of our ChEMU2020 lab. Herein, we describe the resources created for the two tasks, the evaluation methodology adopted, and participants results. We also provide a brief summary of the methods employed by participants of this lab and the results obtained across 46 runs from 11 teams, finding that several submissions achieve substantially better results than the baseline methods prepared by the organizers.

Keywords: Named entity recognition · Event extraction · Information extraction · Chemical reactions · Patent text mining

1 Introduction

Chemical patents represent as an indispensable source information about new discoveries in chemistry. They are usually the first venues where new chemical compounds are disclosed [7,40] and can lead general scientific literature (e.g., journal articles) by up-to 3 years. In addition, chemical patents usually contain much more comprehensive information about the synthesis process of new chemical compounds including their reaction steps and experimental conditions for compound synthesis and mode of action. These details are crucial for the understanding of compound prior art, and provide a means for novelty checking and validation [5,6].

Although the information in chemical patents are of significant research and commercial value, extracting such information is nontrivial, since the large vol-

ume of existing patents and its drastic expansion rate has made manual annotation costly and time-consuming [29]. Natural language processing (NLP) refer to techniques that allow computers to automatically analyze and process natural unstructured language data, and it has enjoyed great success over the past decades [30,44]. In light of this, researchers have been actively exploring the possible application of NLP techniques to patent text mining, so as to alleviate the time-consuming efforts of manual annotation by chemical experts and scale the information extraction process over chemical patents.

The ChEMU (Cheminformatics Elsevier Melbourne University) lab aims to provide a platform for worldwide experts in both NLP and chemistry to develop automated information extraction methods over chemical patents, and to advance the state-of-the-arts in this area. As a first running of ChEMU, our ChEMU2020 lab focuses on extraction of *chemical reactions* from patents [32,14]. Specifically, we provided two information extraction tasks that are crucial steps for chemical reaction extraction. The first task, named entity recognition, requires the identification of essential elements of chemical reactions, such as chemical compounds involved, conditions at which reactions are carried out, and yields of reactions. We go beyond identifying named entities and also require identification of their specific roles in chemical reactions. The second task, event extraction, requires the identification of specific event steps that are performed in a chemical reaction.

In collaboration with chemical domain experts, we have prepared a high-quality annotated data set of 1,500 segments of chemical patent texts specifically targeting these two tasks. The 1,500 segments are sampled from 170 chemical patents, and each segment contains a meaningful chemical reaction. Annotations including entities and event steps are firstly prepared by three chemical experts and then merged to gold-standards.

The ChEMU2020 lab has received considerable interest, attracting 37 registrants from 13 countries including Portugal, Switzerland, Germany, India, Japan, United States, China, and United Kingdom. Specifically, we received 26 runs (1 post-evaluation submission) from 11 teams in Task 1, 10 runs from 5 teams in Task 2, and 10 runs from 4 teams in the task of end-to-end systems (a pipeline combining Task 1 and 2), respectively. Several teams achieved exciting results, outperforming baseline models significantly. In particular, submissions from a team from the company Melax Technologies (from Houston, TX, USA) ranked first in all 3 tasks.

The rest of the paper is structured as follows. We first introduce the corpus we created for use in the lab in Sect. 2. Then we give an overview of the tasks and tracks in Sect. 3, and discuss the evaluation framework used in the lab in Sect. 4. We present the overall evaluation results in Sect. 5 and introduce the participants' approaches in Sect. 6, comparing them in Sect. 7. Conclusions are presented in Sect. 8. Note that this paper is an extension of our previous overview paper [14] and thereby Sect. 2 to 4 here are repeated from that paper; our focus is to provide additional methodological detail.

2 The ChEMU Chemical Reaction Corpus

The annotated corpus prepared for the ChEMU shared task consists of 1,500 patent snippets (text segments) that were sampled from 170 English document patents from the European Patent Office and the United States Patent and Trademark Office. Each snippet contains a meaningful description of a chemical reaction [47].

The corpus was based on information captured in the Reaxys® database.[1] This resource contains details of chemical reactions identified through a mostly manual process of extracting key reaction details from sources including patents and scientific publications, dubbed "excerption" [20].

2.1 Annotation Process

To prepare the gold-standard annotations for the extracted patent snippets, multiple domain experts with rich expert knowledge in chemistry were invited to assist with corpus annotation. A silver-standard annotation set was first generated by mapping the records from the Reaxys database back to the source patents from which the records were originally extracted. This was done by scanning the patent texts for mentions of relevant entities. Since the original records are only linked to the IDs of source patents and do not provide the precise locations of excerpted entities or event steps, these annotations needed to be manually reviewed to produce higher-quality annotations. Two domain experts manually and independently reviewed all patent snippets, correcting location information of the annotations in silver-standard annotations and adding more annotations. Their annotations were then evaluated by measuring their inter-annotator agreement (IAA) [8], and thereafter merged by a third domain expert who acted as an adjudicator, to resolve differences. More details about the quality evaluation over the annotations and the harmonization process will be provided in a more in-depth paper to follow.

We present an example of a patent snippet in Fig. 1. This snippet describes the synthesis of a particular chemical compound, named *N-((5-(hydrazinecarbonyl) pyridin-2-yl)methyl)-1-methyl-N-phenylpiperidine-4-carboxamide*. The synthesis process consists of an ordered sequence of reaction steps: (1) dissolving the chemical compound synthesized in step 3 and hydrazine monohydrate in ethanol; (2) heating the solution under reflux; (3) cooling the solution to room temperature; (4) concentrating the cooled mixture under reduced pressure; (5) purification of the concentrate by column chromatography; and (6) concentration of the purified product to get the title compound.

This shared task aims at extraction of chemical reactions from chemical patents, e.g., extracting the above synthesis steps given the patent snippet in Fig. 1. To achieve this goal, it is crucial for us to first identify the entities that are involved in these reaction steps (e.g., hydrazine monohydrate and ethanol)

[1] https://www.reaxys.com Reaxys® Copyright ©2020 Elsevier Limited except certain content provided by third parties. Reaxys is a trademark of Elsevier Limited.

An example snippet

[Step 4] Synthesis of N-((5-(hydrazinecarbonyl)pyridin-2-yl)methyl)-1-methyl-N-phenylpiperidine-4-carboxamide Methyl 6-((1-methyl-N-phenylpiperidine-4-carboxamido)methyl)nicotinate (0.120 g, 0.327 mmol), synthesized in step 3, and hydrazine monohydrate (0.079 mL, 1.633 mmol) were dissolved in ethanol (10 mL) at room temperature, and the solution was heated under reflux for 12 hours, and then cooled to room temperature to terminate the reaction. The reaction mixture was concentrated under reduced pressure to remove the solvent, and the concentrate was purified by column chromatography (SiO2, 4 g cartridge; methanol/dichloromethane = from 5% to 30%) and concentrated to give the title compound (0.115 g, 95.8%) as a foam solid.

Fig. 1. An example snippet with key focus text highlighted.

and then determine the relations between the involved entities (e.g., hydrazine monohydrate is dissolved in ethanol). Thus, our annotation process consists of two steps: named entity annotations and relation annotations. Next, we describe the two steps of annotations in Sect. 2.2 and Sect. 2.3, respectively.

2.2 Named Entity Annotations

Four categories of entities are annotated over the corpus: (1) **chemical compounds** that are involved in a chemical reaction; (2) **conditions** under which a chemical reaction is carried out; (3) **yields** obtained for the final chemical product; and (4) **example labels** that are associated with reaction specifications.

Ten labels are further defined under the above four categories. We define five different roles that a chemical compound can play within a chemical reaction, corresponding to five labels under this category: STARTING_MATERIAL, REAGENT_CATALYST, REACTION_PRODUCT, SOLVENT, and OTHER_COMPOUND. For example, the chemical compound "ethanol" in Fig. 1 must be annotated with the label "SOLVENT".

We also define two labels under the category of conditions: TIME and TEMPERATURE; and two labels under the category of yields: YIELD_PERCENT and YIELD_OTHER. The definitions of all resultant labels are summarized in Table 1. Interested readers may find more information about the labels in [32] and examples of named entity annotations in the Task 1—NER annotation guidelines [45].

2.3 Relation Annotations

A chemical reaction step typically involves an action and chemical compound(s) on which the action takes effect. We therefore treat the extraction of a reaction

step as a two-stage task: (1) identification of a trigger word that indicates a chemical reaction step; and (2) identification of the relation between a trigger word and chemical compound(s) that is(are) linked to the trigger word. In addition, we observe that it is also crucial for us to link an action to the conditions under which the action is carried out, and resultant yields from the action, in order to fully quantify a reaction step. Thus, annotations in this step are performed to identify the relations between actions (trigger words) and all arguments that are involved in the reaction steps, i.e., chemical compounds, conditions, and yields.

Table 1. Definitions of entity and relation types, i.e., labels, in Task 1 and Task 2.

Label	Definition
Entity Annotations	
STARTING_MATERIAL	A substance that is consumed in the course of a chemical reaction providing atoms to products is considered as starting material.
REAGENT_CATALYST	A reagent is a compound added to a system to cause or help with a chemical reaction.
REACTION_PRODUCT	A product is a substance that is formed during a chemical reaction.
SOLVENT	A solvent is a chemical entity that dissolves a solute resulting in a solution.
OTHER_COMPOUND	Other chemical compounds that are not the products, starting materials, reagents, catalysts and solvents.
TIME	The reaction time of the reaction.
TEMPERATURE	The temperature at which the reaction was carried out.
YIELD_PERCENT	Yield given in percent values.
YIELD_OTHER	Yields provided in other units than %.
EXAMPLE_LABEL	A label associated with a reaction specification.
Relation Annotations	
WORKUP	An event step which is a manipulation required to isolate and purify the product of a chemical reaction.
REACTION_STEP	An event within which starting materials are converted into the product.
Arg1	The relation between an event trigger word and a chemical compound.
ArgM	The relation between an event trigger word and a temperature, time, or yield entity.

Table 2. The annotated entities and trigger words of the snippet example in BRAT standoff format [42].

ID	Entity Type	Offsets	Text Span
T1	TEMPERATURE	313 329	room temperature
T2	REAGENT_CATALYST	231 252	hydrazine monohydrate
T3	REACTION_STEP	281 290	dissolved

Table 3. The annotated relations of the snippet example in BRAT standoff format [42]. Building on the annotations in Table 2, we see that R6 expresses the relation between a compound participating as a reagent (T2) in the T3 "dissolved" reaction step, and R8 captures the temperature (T1) at which that step occurred.

ID	Event Type	Entity 1	Entity 2
R6	Arg1	T3	T2
R8	ArgM	T3	T1

We define two types of trigger words: **WORKUP** which refers to an event step where a chemical compound is isolated/purified, and **REACTION_STEP** which refers to an event step that is involved in the conversion from a starting material to an end product. When labelling event arguments, we adapt semantic argument role labels **Arg1** and **ArgM** from the Proposition Bank [33] to label the relations between the trigger words and other arguments. Specifically, the label Arg1 refers to the relation between an event trigger word and a chemical compound. Here, Arg1 represents argument roles of being causally affected by another participant in the event [16]. ArgM represents adjunct roles with respect to an event, used to label the relation between a trigger word and a temperature, time or yield entity. The definitions of trigger word types and relation types are summarized in Table 1. Detailed annotation guidelines for relation annotation are available online [45].

2.4 Snippet Annotation Format

The gold-standard annotations for the data set were delivered in the BRAT standoff format [42]. For each snippet, two files were delivered: a text file (.txt) containing the original texts in the snippet, and a paired annotation file (.ann) containing all the annotations that have been made for that text, including entities, trigger words, and event steps. Continuing with the above snippet example, we present the formatted annotations for the highlighted sentence in Tables 2 and 3. For ease of presentation, we show the annotated named entities and trigger words in Table 2 and the annotated event steps in Table 3. Specifically, two entities (i.e., T1 and T2) and one trigger word are included in Table 2, and two event steps are included in Table 3.

2.5 Data Partitions

We randomly partitioned the whole data set into three splits for training, development and test purposes, with a ratio of 0.6/0.15/0.25. The training and development sets were released to participants for model development. Note that participants are allowed to use the combination of training and development sets and to use their own partitions to build models. The test set is withheld for use in the formal evaluation. The statistics of the three splits including their number of snippets, total number of sentences, and number of words per snippet, are summarized in Table 4.

To ensure a fair split of data as much as possible, we conduct two statistical tests on the resultant train/dev/test splits. In the first test, we compare the distributions of entity labels (ten classes of entities in Task 1 and two classes of trigger words in Task 2) within train/dev/test sets, to make sure that the three sets of snippets have similar distributions over labels. The distributions are summarize in Table 5, where each cell represents the proportion (e.g., 0.038) of an entity label (e.g., EXAMPLE_LABEL) in the gold annotations of a data split (e.g., Train). The results in Table 5 confirm that the label distributions in the three splits are similar. Only some slight fluctuations ($\leqslant 0.004$) across the three splits are observed for each label.

We further compare the International Patent Classification (IPC) [3] distributions of the training, development and test sets. The IPC information of each patent snippet reflects the application category of the original patent. For example, the IPC code "A61K" represents the category of patents that are for preparations for medical, dental, or toilet purposes. Patents with different IPCs may be written in different ways and may differ in the vocabulary. Thus, they may differ in their linguistic characteristics. For each patent snippet, we extract the primary IPC of its corresponding source patent, and summarize the IPC distributions of the snippets in train/dev/test sets in Table 6.

3 The Tasks

We provide two tasks in ChEMU lab: Task 1—Named Entity Recognition (NER), and Task 2—Event Extraction (EE). We also host a third track where participants can work on building end-to-end systems addressing both tasks jointly.

Table 4. Summary of data set statistics.

Data Split	# snippets	#sentences	# words/snippet
Train	900	5,911	112.16
Dev	225	1,402	104.00
Test	375	2,363	108.63

Table 5. Distributions of entity labels in the training, development, and test sets.

Entity Label	Train	Dev.	Test	Mean
EXAMPLE_LABEL	0.038	0.040	0.037	0.038
OTHER_COMPOUND	0.200	0.198	0.205	0.201
REACTION_PRODUCT	0.088	0.093	0.091	0.091
REAGENT_CATALYST	0.055	0.053	0.053	0.054
SOLVENT	0.049	0.046	0.045	0.047
STARTING_MATERIAL	0.076	0.076	0.075	0.076
TEMPERATURE	0.065	0.064	0.065	0.065
TIME	0.046	0.046	0.048	0.047
YIELD_OTHER	0.046	0.048	0.047	0.047
YIELD_PERCENT	0.041	0.042	0.041	0.041
REACTION_STEP	0.164	0.163	0.160	0.162
WORKUP	0.132	0.132	0.133	0.132

3.1 Task 1: Named Entity Recognition

In order to understand and extract a chemical reaction from natural language texts, the first essential step is to identify the entities that are involved in the chemical reaction. The first task aims to accomplish this step by identifying the ten types of entities described in Sect. 2.2. The task requires the detection of the entity names in patent snippets and the assignment of correct labels to the detected entities (see Table 1). For example, given a detected chemical compound, the task requires the identification of both its text span and its specific type according to the role in which it plays within a chemical reaction description.

3.2 Task 2: Event Extraction

A chemical reaction usually consists of an ordered sequence of event steps that transforms a starting product to an end product, such as the five reaction steps in the synthesis process of the chemical compound described in the example in Figure 1. The event extraction task (Task 2) targets identifying these event steps.

Similarly to conventional event extraction problems [17], Task 2 involves three subtasks: event trigger word detection, event typing and argument prediction. First, it requires the detection of event trigger words and assignment of correct labels for the trigger words. Second, it requires the determination of argument entities that are associated with the trigger words, i.e., which entities identified in Task 1 participate in event or reaction steps. This is done by labelling the connections between event trigger words and their arguments. Given an event trigger word e and a set $\mathcal{S}$ of arguments that participate in e, Task 2 requires the

Table 6. Distributions of International Patent Classifications (IPCs) in the training, development, and test sets. Only dominating IPC groups that take up more than 1 percent of a data split are included in this table.

IPC	Train	Dev.	Test	Mean
A61K	0.277	0.278	0.295	0.283
A61P	0.129	0.134	0.113	0.125
C07C	0.063	0.045	0.060	0.056
C07D	0.439	0.444	0.437	0.440
C07F	0.011	0.009	0.010	0.010
C07K	0.013	0.012	0.008	0.011
C09K	0.012	0.021	0.011	0.015
G03F	0.012	0.019	0.014	0.015
H01L	0.019	0.021	0.019	0.020

creation of $|\mathcal{S}|$ relation entries connecting e to an argument entity in $\mathcal{S}$. Here, $|\mathcal{S}|$ represents the cardinality of the set $\mathcal{S}$. Finally, Task 2 requires the assignment of correct relation type labels (Arg1 or ArgM) to each of the detected relations.

In the track for Task 2, the gold standard entities in snippets are assumed to be known input. While in a real-world use of an event extraction system, gold standard entities would not typically be available, this framework allowed participants to focus on event extraction in isolation of the NER task.

3.3 Task 3: End-to-End Systems

We also hosted a third track which allows participants to develop end-to-end systems that address both tasks simultaneously, i.e., the extraction of reaction events including their constituent entities directly from chemical patent snippets. This is a more realistic scenario for an event extraction system to be applied for large-scale annotation of events.

In the testing stage, participants in this track were provided only with the text of a patent, and were required to identify the named entities defined in Table 1, the trigger words defined in Sect. 3.2, and the event steps involving the entities, that is, the reaction steps. Proposed models in this track were evaluated against the events that they predict for the test snippets, which is the same as in Task 2. However, a major difference between this track and Task 2 is that the gold named entities were not provided but rather had to be predicted by the systems.

3.4 Track overview

We illustrate the workflows of the three tracks in Fig. 2 using as example the sentence highlighted in Fig 1. In Task 1—NER—, participants need to identify

entities that defined in Table 1, e.g., the text span "ethanol" is identified as "SOLVENT". In Task 2—EE—, participants are provided with the three gold standard entities in the sentence. They are required to firstly identify the trigger words and their types (e.g., the text span "dissolved" is identified as "REACTION_STEP") and then identify the relations between the trigger words and the provided entities (e.g., a directed link from "dissolved" to "ethanol" is added and labeled as "ARG1"). In the track of end-to-end systems, participants are only provided with the original text. They are required to identify both the entities and the trigger words, and predict the event steps directly from the text.

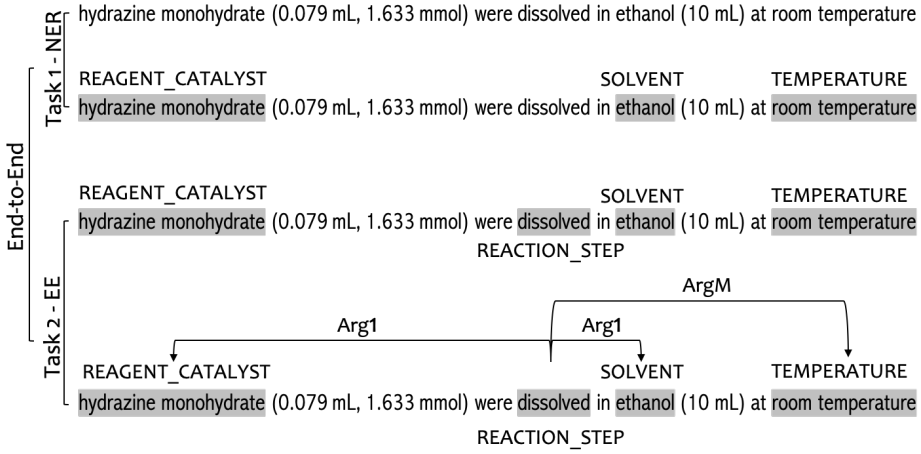

Fig. 2. Illustration of the three tasks. Shaded text spans represents annotated entities or trigger words. Arrows represent relations between entities.

3.5 Organization of Tracks

Training stage. In the training stage, the training and development data sets were released to all participants for model development. To accommodate the needs of participants in different tracks, two different versions of training data, namely Data-NER and Data-EE, were provided. Data-NER was prepared for participants in Task 1, where the gold-standard entities defined in Table 1 were included. Data-EE was prepared for Tasks 2 and 3, where both the gold-standard entities, annotated trigger words and entity relations were included.

Testing stage. Since the gold-standard entities need to be provided to participants in Task 2, the testing stage of Task 2 was delayed until after the testing of Tasks 1 and 3 are completed, in order to prevent any leakage of information. Therefore, the testing stage consists of two phases. In the first phase, the text

(.txt) files of all test snippets were released. Participants in Task 1 are required to use the released patent texts to predict the entities as defined in Table 1. Participants in Task 3 were required to also predict the trigger words and entity relations defined in Sect. 3.2. In the second phase, the gold-standard entities of all test snippets were released. Participants in Task 2 can use the released gold-standard entities, along with the text files released in the first phase, to predict the event steps in test snippets.

Submission website. A submission website has been developed, which allows participants to submit their runs during the testing stage.[2] In addition, the website offers several important functions to facilitate organizing the lab.

First, it hosts the download links for the training, development, and test data sets so that participants can access the data sets conveniently. Second, it allows participants to test the performance (against the development set) of their models before the testing stage starts, which also offers a chance for participants to familiarize themselves with the evaluation tool BRATEval [1] (detailed in Sect. 4). The website also hosts a private leaderboard for each team that ranks all runs submitted by each team, and a public leaderboard that ranks all runs that have been made public by teams.

4 Evaluation Framework

In this section, we describe the evaluation framework of the ChEMU lab. We introduce three baseline algorithms for Task 1, Task 2, and end-to-end systems, respectively.

4.1 Evaluation Methods

We use BRATEval [1] to evaluate all the runs that we receive. Three metrics are used to evaluate the performance of all the submissions for Task 1: Precision, Recall, and F_1-score. Specifically, given a predicted entity and a ground-truth entity, we treat the two entities as a match if (1) the types associated with the two entities match; and (2) their text spans match. The overall Precision, Recall, and F_1-score are computed by micro-averaging all instances (entities).

In addition, we exploit two different matching criteria, exact-match and relaxed-match, when comparing the texts spans of two entities. Here, the exact-match criterion means that we consider that the text span of an entity matches with that of another entity if both the starting and the end offsets of their spans match. The relaxed-match criterion means that we consider that the text span of one entity matches with that of another entity as long as their text spans overlap.

The submissions for Task 2 and end-to-end systems are evaluated using Precision, Recall, and F_1-score by comparing the predicted events and gold standard

[2] http://chemu.eng.unimelb.edu.au/

events. We consider two events as a match if (1) their trigger words and event types are the same; and (2) the entities involved in the two events match. Here, we follow the method in Task 1 to test whether two entities match. This means that the matching criteria of exact-match and relaxed-match are also applied in the evaluation of Task 2 and of end-to-end systems. Note that the relaxed-match will only be applied when matching the spans of two entities; it does not relax the requirement that the entity type of predicted and ground truth entities must agree. Since Task 2 provides gold entities but not event triggers with their ground truth spans, the relaxed-match only reflects the accuracy of spans of predicted trigger words.

To somewhat accommodate a relaxed form of entity type matching, we also evaluate submissions in Task 1—NER using a set of high-level labels shown in the hierarchical structure of entity classes in Fig. 3. The higher-level labels used are highlighted in grey. In this set of evaluations, given a predicted entity and a ground-truth entity, we consider that their labels match as long as their corresponding high-level labels match. For example, suppose we get as predicted entity "STARTING_MATERIAL, [335, 351), boron tribromide" while the (correct) ground-truth entity instead reads "REAGENT_CATALYST, [335, 351), boron tribromide", where each entity is presented in the form of "TYPE, SPAN, TEXT". In the evaluation framework described earlier this example will be counted as a mismatch. However, in this additional set of entity type relaxed evaluations we consider the two entities as a match, since both labels "STARTING_MATERIAL" and "REAGENT_CATALYST" specialize their parent label "COMPOUND".

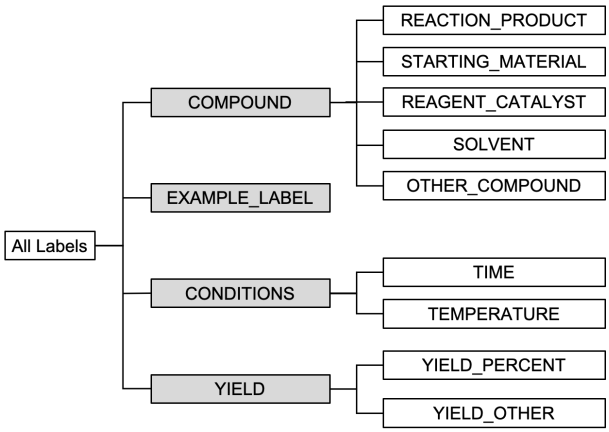

Fig. 3. Illustration of the hierarchical NER class structure used in evaluation.

4.2 Baselines

We released one baseline method for each task as a benchmark method. Specifically, the baseline for Task 1 is based on retraining **BANNER** [21] on the training and development data; the baseline for Task 2 is a co-occurrence method; and the baseline for end-to-end systems is a two-stage algorithm that first uses BANNER to identify entities in the input and then uses the co-occurrence method to extract events.

BANNER. BANNER is a named entity recognition tool for bio-medical data. In this baseline, we first use the GENIA Sentence Splitter (GeniaSS) [38] to split input texts into separate sentences. The resulting sentences are then fed into BANNER, which predicts the named entities using three steps, namely tokenization, feature generation, and entity labelling. A simple tokenizer is used to break sentences into either a contiguous block of letters and/or digits or a single punctuation mark. BANNER uses a conditional random field (CRF) implementation derived from the MALLET toolkit[3] for feature generation and token labelling. The set of machine learning features used consist primarily of orthographic, morphological and shallow syntax features.

Co-occurrence Method. This method first creates a dictionary D_e for the observed trigger words and their corresponding types from the training and development sets. For example, if a word "added" is annotated as a trigger word with the label of "WORKUP" in the training set, we add an entry $\langle \text{added}, \text{WORKUP} \rangle$ to D_e. In the case where the same word has been observed to appear as both types of "WORKUP" and "REACTION_STEP", we only keep as entry in D its most frequent label. The method also creates an event dictionary D_r for the observed event types in the training and development sets. For example, if an event $\langle \text{ARG1}, \text{E1}, \text{E2} \rangle$ is observed where "E1" corresponds to trigger word "added" of type "WORKUP" and "E2" corresponds to entity "water" of type "OTHER_COMPOUND", we add an entry $\langle \text{ARG1}, \text{WORKUP}, \text{OTHER\_COMPOUND} \rangle$ to D_r.

To predict events, this method first identifies all trigger words in the test set using D_e. It then extracts two events $\langle \text{ARG1}, \text{T1}, \text{T2} \rangle$ and $\langle \text{ARGM}, \text{T1}, \text{T2} \rangle$ for a trigger word "E1" and an entity "E2" if (1) they co-occur in the same sentence; and (2) the relation type $\langle \text{ARGx}, \text{T1}, \text{T2} \rangle$ is included in D_r. Here, "ARGx" can be "ARG1" or "ARGM", and "T1" and "T2" are the entity types of "E1" and "E2" respectively.

BANNER + Co-occurrence Method. The above two baselines are combined to form a two-stage method for end-to-end systems. This baseline first uses BANNER to identify all the entities in Task 1. Then it utilizes the co-occurrence method to predict events, except that gold standard entities are replaced with the entities predicted by BANNER in the first stage.

[3] http://mallet.cs.umass.edu/

5 Results

In total, 39 teams registered for the ChEMU shared task, of which 36 teams registered for Task 1, 31 teams registered for Task 2, and 28 teams registered for both tasks. The 39 teams are spread across 13 different countries, from both the academic and industry research communities. In this section, we report the results of all the runs that we received for each task.

5.1 Task 1—Named Entity Recognition

Task 1 received considerable interest with the submission of 25 runs from 11 teams. The 11 teams include 1 team from Germany (OntoChem), 3 teams from India (AUKBC, SSN_NLP and JU_INDIA), 1 team from Switzerland (BiTeM), 1 team from Portugal (Lasige_BioTM), 1 team from Russia (KFU_NLP), 1 team from the United Kingdom (NextMove Software/Minesoft), 2 teams from the United States of America (Melaxtech and NLP@VCU), and 1 team from Vietnam (VinAI). We evaluate the performance of all 25 runs, comparing their predicted entities with the ground-truth entities of the patent snippets in the test set. We report the performances of all runs under both matching criteria in terms of three metrics, namely Precision, Recall, and F_1-score.

We report the overall performance of all runs in Table 7. The baseline of Task 1 achieves 0.8893 in F_1-score under exact match. Nine runs outperform the baseline in terms of F_1-score under exact match. The best run was submitted by team Melaxtech, achieving a high F_1-score of 0.9570. There were sixteen runs with an F_1-score greater than 0.90 under relaxed-match. However, under exact-match, only seven runs surpassed 0.90 in F_1-score. This difference between exact-match and relaxed-match may be related to the long text spans of chemical compounds, which is one of the main challenges in NER tasks in the domain of chemical documents.

Next, we evaluate the performance of all 25 runs using the high-level labels in Fig. 3 (highlighted in grey). We report the performances of all runs in terms of Precision, Recall, and F_1-score in Table 8.

5.2 Task 2—Event Extraction

We received 10 runs from five teams. Specifically, the five teams include 1 team from Portugal (Lasige_BioTM), 1 team from Turkey (BOUN_REX), 1 team from the United Kingdom (NextMove Software/Minesoft) and 2 teams from the United States of America (Melaxtech and NLP@VCU). We evaluate all runs using the metrics Precision, Recall, and F_1-score. Again, we utilize the two matching criteria, namely exact-match and relaxed-match, when comparing the trigger words in the submitted runs and ground-truth data.

The overall performance of each run is summarized in Table 9.[4] The baseline (co-occurrence method) scored relatively high in Recall, i.e, 0.8861. This was

[4] The run that we received from team Lasige_BioTM is not included in the table due to a technical issue found in this run.

Table 7. Overall performance of all runs in Task 1—Named Entity Recognition. Here, P, R, and F represents the Precision, Recall, and F_1-score, respectively. For each metric, we highlight the best result in **bold** and the second best result in *italic*. The results are ordered by their performance in terms of F_1-score under exact-match. *This run was received after evaluation phase and thus was not included in official results.

Run	Exact-Match			Relaxed-Match		
	P	R	F	P	R	F
Melaxtech-run1	0.9571	**0.9570**	**0.9570**	0.9690	*0.9687*	*0.9688*
Melaxtech-run2	**0.9587**	*0.9529*	*0.9558*	0.9697	0.9637	0.9667
Melaxtech-run3	*0.9572*	0.9510	0.9541	0.9688	0.9624	0.9656
VinAI-run2*	0.9538	0.9504	0.9521	**0.9737**	**0.9716**	**0.9726**
VinAI-run1	0.9462	0.9405	0.9433	*0.9707*	0.9661	0.9684
Lasige_BioTM-run1	0.9327	0.9457	0.9392	0.9590	0.9671	0.9630
BiTeM-run3	0.9378	0.9087	0.9230	0.9692	0.9558	0.9624
BiTeM-run2	0.9083	0.9114	0.9098	0.9510	0.9684	0.9596
NextMove/Minesoft-run1	0.9042	0.8924	0.8983	0.9301	0.9181	0.9240
NextMove/Minesoft-run2	0.9037	0.8918	0.8977	0.9294	0.9178	0.9236
Baseline	0.9071	0.8723	0.8893	0.9219	0.8893	0.9053
NLP@VCU-run1	0.8747	0.8570	0.8658	0.9524	0.9513	0.9518
KFU_NLP-run1	0.8930	0.8386	0.8649	0.9701	0.9255	0.9473
NLP@VCU-run2	0.8705	0.8502	0.8602	0.9490	0.9446	0.9468
NLP@VCU-run3	0.8665	0.8514	0.8589	0.9486	0.9528	0.9507
KFU_NLP-run2	0.8579	0.8329	0.8452	0.9690	0.9395	0.9540
NextMove/Minesoft-run3	0.8281	0.8083	0.8181	0.8543	0.8350	0.8445
KFU_NLP-run3	0.8197	0.8027	0.8111	0.9579	0.9350	0.9463
BiTeM-run1	0.8330	0.7799	0.8056	0.8882	0.8492	0.8683
OntoChem-run1	0.7927	0.5983	0.6819	0.8441	0.6364	0.7257
AUKBC-run1	0.6763	0.4074	0.5085	0.8793	0.5334	0.6640
AUKBC-run2	0.4895	0.1913	0.2751	0.6686	0.2619	0.3764
SSN_NLP-run1	0.2923	0.1911	0.2311	0.8633	0.4930	0.6276
SSN_NLP-run2	0.2908	0.1911	0.2307	0.8595	0.4932	0.6267
JU_INDIA-run1	0.1411	0.0824	0.1041	0.2522	0.1470	0.1857
JU_INDIA-run2	0.0322	0.0151	0.0206	0.1513	0.0710	0.0966
JU_INDIA-run3	0.0322	0.0151	0.0206	0.1513	0.0710	0.0966

Table 8. Overall performance of all runs in Task 1—Named Entity Recognition where the set of high-level labels in Fig. 3 is used. Here, P, R, and F represents the Precision, Recall, and F_1-score, respectively. For each metric, we highlight the best result in **bold** and the second best result in *italic*. The results are ordered by their performance in terms of F_1-score under exact-match. *This run was received after evaluation phase and thus was not included in official results.

Run	Exact-Match			Relaxed-Match		
	P	R	F	P	R	F
Melaxtech-run1	*0.9774*	**0.9774**	**0.9774**	0.9906	0.9901	0.9903
Melaxtech-run2	**0.9789**	*0.9732*	*0.9760*	0.9910	0.9849	0.9879
Melaxtech-run3	0.9775	0.9714	0.9744	0.9905	0.9838	0.9871
VinAI-run2*	0.9704	0.9670	0.9687	**0.9920**	0.9901	*0.9911*
Lasige_BioTM-run1	0.9571	0.9706	0.9638	0.9886	*0.9943*	**0.9915**
VinAI-run1	0.9635	0.9579	0.9607	0.9899	0.9854	0.9877
Baseline	0.9657	0.9288	0.9469	0.9861	0.9519	0.9687
BiTeM-run1	0.9573	0.9277	0.9423	0.9907	0.9770	0.9838
NextMove/Minesoft-run2	0.9460	0.9330	0.9394	0.9773	0.9611	0.9691
NextMove/Minesoft-run1	0.9458	0.9330	0.9393	0.9773	0.9610	0.9691
BiTeM-run2	0.9323	0.9357	0.9340	0.9845	**0.9962**	*0.9903*
NextMove/Minesoft-run3	0.9201	0.8970	0.9084	0.9571	0.9308	0.9438
NLP@VCU-run1	0.9016	0.8835	0.8925	0.9855	0.9814	0.9834
NLP@VCU-run2	0.9007	0.8799	0.8902	0.9882	0.9798	0.9840
NLP@VCU-run3	0.8960	0.8805	0.8882	0.9858	0.9869	0.9863
KFU_NLP-run1	0.9125	0.8570	0.8839	*0.9911*	0.9465	0.9683
BiTeM-run3	0.9073	0.8496	0.8775	0.9894	0.9355	0.9617
KFU_NLP-run2	0.8735	0.8481	0.8606	0.988	0.9569	0.9722
KFU_NLP-run3	0.8332	0.8160	0.8245	0.9789	0.9516	0.9651
OntoChem-run1	0.9029	0.6796	0.7755	0.9611	0.7226	0.8249
AUKBC-run1	0.7542	0.4544	0.5671	0.9833	0.5977	0.7435
AUKBC-run2	0.6605	0.2581	0.3712	0.9290	0.3612	0.5201
SSN_NLP-run2	0.3174	0.2084	0.2516	0.9491	0.5324	0.6822
SSN_NLP-run1	0.3179	0.2076	0.2512	0.9505	0.5304	0.6808
JU_INDIA-run1	0.2019	0.1180	0.1489	0.5790	0.3228	0.4145
JU_INDIA-run2	0.0557	0.0262	0.0357	0.4780	0.2149	0.2965
JU_INDIA-run3	0.0557	0.0262	0.0357	0.4780	0.2149	0.2965

expected, since the co-occurrence method aggressively extracts all possible events within a sentence. However, the F_1-score was low due to its low Precision score. Here, all runs outperform the baseline in terms of F_1-score under exact-match. Melaxtech ranks first among all official runs in this task, with an F_1-score of 0.9536.

Table 9. Overall performance of all runs in Task 2—Event Extraction. Here, P, R, and F represent the Precision, Recall, and F_1-score, respectively. For each metric, we highlight the best result in **bold** and the second best result in *italics*. The results are ordered by their performance in terms of F_1-score under exact-match.

Run	Exact-Match			Relaxed-Match		
	P	R	F	P	R	F
Melaxtech-run1	*0.9568*	**0.9504**	**0.9536**	*0.9580*	**0.9516**	**0.9548**
Melaxtech-run2	**0.9619**	0.9402	*0.9509*	**0.9632**	0.9414	*0.9522*
Melaxtech-run3	0.9522	*0.9437*	0.9479	0.9534	*0.9449*	0.9491
NextMove/Minesoft-run1	0.9441	0.8556	0.8977	0.9441	0.8556	0.8977
NextMove/Minesoft-run2	0.8746	0.7816	0.8255	0.8909	0.7983	0.8420
BOUN_REX-run1	0.7610	0.6893	0.7234	0.7610	0.6893	0.7234
NLP@VCU-run1	0.8056	0.5449	0.6501	0.8059	0.5451	0.6503
NLP@VCU-run2	0.5120	0.7153	0.5968	0.5125	0.7160	0.5974
NLP@VCU-run3	0.5085	0.7126	0.5935	0.5090	0.7133	0.5941
Baseline	0.2431	0.8861	0.3815	0.2431	0.8863	0.3816

5.3 End-to-end Systems

We received 10 end-to-end system runs from four teams. The four teams include The four teams include 1 team from Turkey (BOUN_REX), 1 team from the United Kingdom (NextMove Software/Minesoft) and 2 teams from the United States of America (Melaxtech and NLP@VCU).

The overall performance of all runs is summarized in Table 10 in terms of Precision, Recall, and F_1-score under both exact-match and relaxed-match.[5] Since gold entities are not provided in this task, the average performance of the runs in this task are slightly lower than those in Task 2. Note that the Recall scores of most runs are substantially lower than their Precision scores. This may reveal that the task of identifying a relation from a chemical patent is harder

[5] The run that we received from the Lasige_BioTM team is not included in the table as there was a technical issue in this run. Two runs from Melaxtech, Melaxtech-run2 and Melaxtech-run3, had very low performance, due to an error in their data pre-processing step.

Table 10. Overall performance of all runs in end-to-end systems. Here, P, R, and F represent the Precision, Recall, and F_1-score, respectively. For each metric, we highlight the best result in **bold** and the second best result in *italics*. The results are ordered by their performance in terms of F_1-score under exact-match.

Run	Exact-Match			Relaxed-Match		
	P	R	F	P	R	F
Melaxtech-run1	**0.9201**	**0.9147**	**0.9174**	**0.9319**	**0.9261**	**0.9290**
NextMove/Minesoft-run1	*0.8492*	*0.7609*	*0.8026*	*0.8663*	*0.7777*	*0.8196*
NextMove/Minesoft-run2	0.8486	0.7602	0.8020	0.8653	0.7771	0.8188
NextMove/Minesoft-run3	0.8061	0.7207	0.7610	0.8228	0.7371	0.7776
OntoChem-run1	0.7971	0.3777	0.5126	0.8407	0.3984	0.5406
OntoChem-run2	0.7971	0.3777	0.5126	0.8407	0.3984	0.5406
OntoChem-run3	0.7971	0.3777	0.5126	0.8407	0.3984	0.5406
Baseline	0.2104	0.7329	0.3270	0.2135	0.7445	0.3319
Melaxtech-run2	0.2394	0.2647	0.2514	0.2429	0.2687	0.2552
Melaxtech-run3	0.2383	0.2642	0.2506	0.2421	0.2684	0.2545

than the task of typing an identified relation. The first run from Melaxtech team ranks best among all runs received for this task.

6 Overview of Participants' Approaches

We received 8 paper submissions from participating teams, namely BiTeM, VinAI, BOUN-REX, NextMove/Minesoft, NLP@VCU, AU-KBC, LasigBioTM, and MelaxTech. In this section, we present an overview of the approaches proposed by these teams. We start by introducing the approach of each team first. Then we discuss the differences between these approaches.

6.1 BiTeM

To tackle the complexities of chemical patent narratives, the BiTeM team explored the power of ensemble of deep neural language models based on transformer architectures to extract information in chemical patents [10]. Using a majority vote strategy [9], their approach combined end-to-end architectures, including Bidirectional Encoder Representations from Transformers (BERT) models (including both base/large and cased/uncased) [12], the ChemBERTa model[6], and a model based on Convolutional Neural Network (CNN) [22] fed with contextualized embedding vectors provided by BERT model. To learn to classify chemical entities in patent passages, the language models were fine-tuned to

[6] https://github.com/seyonechithrananda/bert-loves-chemistry

categorize tokens using training examples of the ChEMU NER task. The best model proposed by BiTeM – an ensemble of BERT-base cased and uncased, and a CNN – achieved 92.3% of exact F_1-score and 96.24% of relaxed F_1-score in the test phase, outperforming the exact F_1-score of the best individual model (BERT-base cased) by 1.3% and the challenge's baseline by 3.4%. The results of BiTeM team show that ensemble of contextualized language models could be used to effectively detect chemical entities in patent narratives.

6.2 VinAI

Following [48], the VinAI system employed the well-known BiLSTM-CNN-CRF model [26] with additional contextualized word embeddings. In particular, given an input sequence of words, VinAI represented each word token by concatenating its corresponding pre-trained word embedding, CNN-based character-level word embedding [26] and contextualized word embedding. Here, VinAI used the pre-trained word embeddings released by [48], which are trained on a corpus of 84K chemical patents using the Word2Vec skip-gram model [28]. Also, VinAI employed the contextualized word embeddings generated by a pre-trained ELMo language model [35] which is trained using the same corpus of 84K chemical patents [48].[7] Then the concatenated word representations are fed into a BiLSTM encoder to extract latent feature vectors for input words. Each latent feature vector is then linearly transformed before being fed into a linear-chain CRF layer for NER label prediction [19]. VinAI achieved very high performance, officially ranking second with regards to both exact- and relaxed-match F_1-scores at 94.33% and 96.84%, respectively. In a post-evaluation phase, fixing a mapping bug which converted the column-based format into the brat standoff format helped VinAI to obtain higher results: an exact-match F_1-score at 95.21% and especially a relaxed-match F_1-score at 97.26%, thus achieving the highest relaxed-match F_1-score compared with all other participating systems.

6.3 BOUN-REX

The BOUN-REX system addressed the event extraction task with two steps: the detection of trigger word and its trigger type, and identification of the event type. A pre-trained transformer-based model, BioBERT, was used for the detection of trigger words (and the exact types of the detected trigger words) whereas the event type is determined using a rule-based method. Specifically, to pre-process the dataset, the documents were first split into sentences via GENIA Sentence Splitter. After constructing sentence-entity pairs for each entity, events and trigger words were predicted from the given sentence-entity pairs. Start and end markers were also introduced for each entity to indicate position of an entity in a sentence. A pre-trained transformer-based architecture was constructed as a base model to extract a fixed-length relation representation and token representations from an input sentence with entity markers. The fixed-length relation

[7] https://github.com/zenanz/ChemPatentEmbeddings

representation was utilized to detect the type of the trigger word. In addition, a trigger word span model was also constructed to predict the probabilities of start and end markers of trigger words with the token representations. The system trained using an AdamW optimizer, achieving the best performance of an F_1 score at 0.7234 using exact match.

6.4 NextMove Software/Minesoft

Lowe and Mayfield [24] used an approach utilizing grammars both for recognizing entities and for determining the relationships between entities. The toolkit LeadMine was first used to recognized chemicals and physical quantities, which is achieved by employing efficient matching against extensive grammars that describe these entity types. These entities were used as the highest priority tagger in an enhanced version of ChemicalTagger, with ChemicalTagger's default rule based tokenization being adjusted such that each LeadMine entity was a single token. Remaining tokens were assigned tags from pattern matches, or failing that assigned a part of speech tag. The pattern matches notably are how the reaction action trigger words are detected. An Antlr grammar arranges the tagged tokens into a parse tree which groups at various levels of granularity, e.g. all tokens for a particular reaction action will be grouped. The parse tree is used to determine which chemicals are solvents or involved in workup actions. The chemical structures (determined from the names), is used to assign chemical role information, both by inspection of the individual compounds and through whole reaction analysis techniques like NameRxn and atom-atom mapping. From analysis of the whole reaction the stoichiometry of the reaction is determined, hence distinguishing catalysts from starting materials.

6.5 VLP@VCU

VLP@VCU team participated in two tasks: Task 1—NER and Task 2–EE. For Task 1, the VLP@VCU team identified the named entities using BiLSTM units with a Conditational Random Graph (CRF) output layer. The inputs to this model are pre-trained word embeddings [32] in combination with character embeddings. These embeddings are concatenated and then passed through the BiLSTM network. The VLP@VCU system achieved an overall performance with a precision, recall and F_1-score at 0.87, 0.86, and 0.87 in terms of exact match, and a precision, recall and F_1-score of 0.95, 0.99, and 0.97 in terms of relaxed match.

For Task 2—EE, the VLP@VCU team explored two methods to identify the chemical arguments between the trigger words and the entities. First, a rule-based method was explored, which uses a breadth-first search to find the closest occurrence of the trigger word on either side of the entity. Second, a CNN-based model was explored. This model performs a binary classification to identify whether there is a relation or not for each *Trigger word-Entity pair*. The sentence containing the *Trigger word-Entity pair* is first extracted and then divided into five segments, where each segment is represented by a $k \times N$ matrix. Here,

k represents the latent dimensionality of the pre-trained word embeddings [32] and N is the number of words in the segment. A separate convolution unit is constructed for each segment, the outputs of which are then flattened, concatenated, and fed into the fully connected feedforward layer. Finally, the output of the fully connected feedforward layer is fed into a softmax layer, which performs the final classification. This CNN-based method obtained higher performance with a precision, recall and F_1-score of 0.80, 0.54 and 0.65, respectively.

6.6 AU-KBC

The AU-KBC team submitted two systems that were developed with two different Machine Learning (ML) techniques: CRFs and Artificial Neural Networks (ANNs). A two-stage pre-processing was done on the training and development data sets: (1) a formatting stage that consists of three steps, i.e., sentence splitting, tokenization, and conversion of data format to column; and (2) a data annotation stage, where the data is annotated for syntactic information, including Part-of-Speech (PoS) and Phrase Chunk information (noun phrase, verb phrase). To extract the PoS and chunk information, an open source tool, fnTBL [31], is used. Three types of features were used for training: (a) word-level features, (b) grammatical features, and (c) functional terms features. Specifically, word-level features include orthographical features (e.g., capitalization, Greek words, and combination of digits, symbols, and words) and morphological features (e.g., common prefixes and suffixes of chemical entities). Grammatical features include word, Part-of-Speech (PoS), chunks and combination of PoS and chunks. Functional terms were used to help identify the biological named entities and assign them with correct labels. After extraction of these linguistic features, two models based on CRF and ANN are built to address Task 1. Note that the two models only utilized the training data provided in the task and did not rely on any external resources or leverage pre-trained language models. Specifically, the CRF++ tool [2] was used for developing the CRF model and the ANN model was implemented using the scikit python package. The ANN model is a Multi-Layer Perceptron (MLP), where ReLU activation function was used. The stochastic gradient Adam optimizer was used for optimizing weights of the ANN model. We obtained an F_1-score of 0.6640 using CRFs and F_1-score of 0.3764 using ANN.

6.7 LasigBioTM

To address Task 1, the LasigBioTM team fine-tuned the BioBERT NER model in the train set plus half of the development set (Note that the data was converted to the IOB2 format), and applied the fine-tuned model into the second half of the development set for recognizing and locating named entities. The team also developed a module to handle the BioBERT output and to generate the annotation files in the BRAT format and then submitted them to the competition page for evaluation, obtaining an F_1-score of 0.9524 using the exact matching criterion and a F_1-score of 0.9904 using the relaxed matching criterion on development

data. In the testing phase, the team fine-tuned the model again, but using all the documents belonging to the train and the development sets. For Task 2, the team considered the BioBERT NER model jointly with the BioBERT RE model. They followed a similar approach as for Task 1 to detect the trigger words. To further extract the relations between triggers and entities, the team performed sentence segmentation of the train and the development sets and, if a trigger word and an entity were present in a given sentence, a relation was assumed to exist between them if it was referred in the respective annotation file. The BioBERT RE model was also fine-tuned using the sentences of the train and the development sets.

6.8 MelaxTech

The MelaxTech system is a hybrid combination of deep learning models and pattern-based rules for this task. For deep learning, a language model of patents with chemical reactions was first built. Specifically, the BioBERT [23], a pre-trained biomedical language model (a bidirectional encoder representation for biomedical text), was used as the basis for training a language model of patents. Based on BERT [12], a language model pre-trained on large scale open text, BioBERT was further refined using the biomedical literature in PubMed and PMC. For this study, BioBERT was retrained on patent data to generate a new language model of Patent_BioBERT. For the NER subtask, Patent_BioBERT was fine-tuned using the Bi-LSTM-CRF (Bi-directional Long-Short-Term-Memory Conditional-Random-Field) algorithm [26]. Next, several rules based on observed patterns in the training data were used in a post-processing step. For example, rules were defined to differentiate STARTING_MATERIAL and OTHER_COMPOUND based on the relative positions and total number of EXAMPLE_LABEL occurrences. For the event extraction subtask, the event triggers were first identified as named entities together with other semantic roles in chemical reaction, using the same approach as in the NER subtask. Next, a binary classifier was built by fine-tuning Patent_BioBERT to recognize relations between event triggers and semantic roles in the same sentence. Some event triggers and their linked semantic roles were present in different sentences, or different clauses in long complex sentences. Their relations were not identified using the deep learning-based model. Therefore, post-processing rules were designed based on patterns observed in the training data, and applied to recover some of these false negative relations. The proposed approaches demonstrated promising results, which achieved top ranks in both subtasks, with the best F_1-score of 0.957 for entity recognition and the best F_1-score of 0.9536 for event extraction.

7 Discussion

Different approaches were explored by the participating teams. In Table 11, we summarize the key strategies in terms of three aspects: tokenization method, token representations, and core model architecture.

For teams who participated in Tasks 2 and 3, a common two-step strategy was adopted for relation extraction: (1) identify trigger words; and (2) extract the relation between identified trigger words and entities. The first step is essentially an NER task, and the second step is a relation extraction task. As such, NER models were used by all these teams for Tasks 2 and 3 as well as by the teams participating in Task 1. Therefore, in what follows, we first discuss and compare the approaches of all teams without considering the target tasks, subsequently considering relation extraction approaches.

Table 11. Summary of participants' approaches. [10]: BiTeM; [11]: VinAI; [18]: BOUNREX; [24]: NextMove/Minesoft; [27]: NLP@VCU; [34]: AU-KBC; [37]: LasigBioTM; and [46]: MelaxTech.

Characteristics	[10]	[11]	[18]	[24]	[27]	[34]	[37]	[46]
Tokenization								
Rule-based	☑	☑	☑	☑	☑	☑	☑	☑
Dictionary-based				☑				
Subword-based	☑		☑				☑	☑
Chemistry domain-specific		☑		☑				
Representation								
Embeddings								
Character-level		☑			☑			
Pre-trained	☑	☑	☑		☑		☑	☑
Chemistry domain-specific		☑	☑		☑		☑	☑
Features								
PoS				☑		☑		
Phrase				☑		☑		
Model Architecture								
Transformer	☑		☑				☑	☑
Bi-LSTM		☑			☑			
CNN	☑	☑			☑			
MLP						☑		
CRF	☑	☑			☑	☑		
FSM				☑				
Rule based			☑	☑	☑			☑

7.1 Tokenization

Tokenization is an important data pre-processing step that splits input texts into words/subwords, i.e., tokens. We identify three general types of tokenization methods used by participants: (1) rule based tokenization; (2) dictionary based tokenization; and (3) subword based tokenization. Specifically, rule based tokenization applies pre-defined rules to split texts into tokens. The rules applied can be as simple as "white-space tokenization", but can also be a complex mixture of a set of carefully designed rules (e.g., based on language-specific grammar rules and common prefixes). Dictionary based tokenization requires the construction of a vocabulary and the text splitting is performed by matching the input text with the existing tokens in the constructed vocabulary. Subword tokenization allows a token to be a sub-string of a word, i.e., subword units. It relies on the principle that most common words should be left as is, but rare words should be decomposed in meaningful subword units. Popular subword tokenization methods include WordPiece [39] and Byte Pair Encoding (BPE) [36]. For each participating team, we consider whether their approach belong to one or multiple of the three categories, and summarize our findings in Table 11. Finally, we also indicate whether their tokenization methods consider domain-specific knowledge in Table 11.

Four teams utilized tokenization methods that are purely rule-based. Specifically, VinAI used Oscar4 tokenizer [15]. This tokenizer is particularly designed for chemical texts, and is made up by a set of pattern matching rules (e.g., prefix matching) that are designed based on domain knowledge from chemical experts. NLP@VCU used Spacy tokenizer [4], which consists of a collection of complex normalization and segmentation logic and has been proven to work well with general English corpus. NextMove/Minesoft used a combination of Oscar4 and LeadMine [25] tokenzier. LeadMine was first run on untokenized text to identify entities using auxiliary grammars or dictionaries. Oscar4 was then used for general tokenization but is adjusted so that each entity recognized by LeadMine corresponds to exactly one token. Four teams, BiTeM, BOUN-REX, LasigBioTM, and MelaxTech, chose to leverage the pre-trained model BERT (or variants of BERT) to address our tasks, and thus, the four teams used the subword-based tokenizer, WordPiece, that is built-in within BERT. BOUN-REX, LasigBioTM and MelaxTech used BioBERT model which is language model pre-trained on biomedical texts. Since this model is a continual training based on the original BERT model, the vocabulary used in BioBERT does not differ from BERT, i.e., domain-specific tokenization is not used. However, since MelaxTech performed a pre-tokenization using a toolkit CLAMP [41], we consider their approach as domain-specific, since CLAMP is tailored for clinical texts. BiTeM used the model ChemBERTa that is pre-trained on ZINC corpus. It is unclear yet whether the tokenization is domain-specific due to the lack of documentation of ChemBERTa. Finally, since WordPiece needs an extra pre-tokenization step, we consider it as a hybrid of rule-based and subword-based method.

7.2 Representations

When transforming tokens into machine-readable representations, two types of methods are used: (1) feature extraction that represents tokens with their linguistic characteristics such as word-level features (e.g., morphological features) and grammatical features (e.g., PoS tags); and (2) embedding methods in which token representations are randomly initialized as numerical vectors (or initialized from pre-trained embeddings) and then learned (or fine-tuned) from provided training data. Two teams, NextMove/Minesoft and AU-KBC adopted the first strategy and the other teams adopted the second strategy. Among the teams that used embeddings to represent tokens, two teams, VinAI and VLP@VCU further added character-level embeddings to their systems. All of these six teams used pre-trained embeddings, and five teams used embeddings that are pre-trained for related domains: VinAI and NLP@VCU used the embeddings that are pre-trained on chemical patents [48], BOUN-REX and LasigBioTM used the embeddings from BioBERT model that are pre-trained on PubMed corpus. MexlaxTech also used embeddings from BioBERT, but they further tuned the embeddings using the patent documents released in the test phase.

7.3 Model Architecture

Various architectures were employed by participating teams. Four teams, BiTeM, BOUN-REX, LasigBioTM, and MelaxTech developed their systems based on transformers [43]. BiTem submitted an additional run using an ensemble of a Transformer-based model and a CNN-based model. They also had a third run that is built based on CRF. The other two teams MelaxTech and BOUN-REX added rule-based techniques into their systems. MelaxTech added several pattern-matching rules in their post-processing step. BOUN-REX focused on Task 2 and their system used rule based methods to determine the event type of each detected event. Two teams, VinAI and NLP@VCU, used the architecture of BiLSTM-CNN-CRF for Task 1. NLP@VCU also participated in Task 2 and they proposed two systems based on rules, and CNN architecture, respectively. NextMove/Minesoft utilized Chemical Tagger [13], a model based on Finite State Machine (FSM), and a set of comprehensive rules are applied to generate predictions. AU-KBC proposed two systems for Task 1, based on multi-layer perceptron and CRF, respectively.

7.4 Approaches to relation extraction

Four of the above teams participated in Task 2 or Task 3. As mentioned before, these teams utilized their NER models for trigger word detection. Thus, here, we only discuss their approaches for relation extraction assuming that the trigger words and entities are known.

NextMove/Minesoft again made use of ChemicalTagger for event extraction. ChemicalTagger is able to recognize WORKUP and REACTION_STEP words, thus, assignment of relationships were achieved by associating all entities in a

ChemicalTagger action phrase with the trigger word responsible for the action phrase. A set of post-processing rules were also applied to enhance the accuracy of ChemicalTagger.

LasigBioTM, NLP@VCU, and MelaxTech formulated the task of relation extraction as a binary classification problem. That is, given each candidate pair of trigger word and named entity that co-locate within an input sentence, the goal of the task is to determine whether the candidate pair of entities are related or not.

LasigBioTM developed a BioBERT-based model to accomplish this classification. The input of BioBERT is the sentence containing the candidate pair but the trigger word and named entity of the candidate pair were replaced with the tags "@TRIGGER$" and "@LABEL$", respectively. The output of BioBERT is modified as a binary classification layer which aims to predict the existence of relation for the candidate pair.

NLP@VCU proposed two systems for relation extraction. Their first system is a rule-based system. Given a named entity, a relation is extracted between the named entity and its nearest trigger word. Their second system is developed based on CNNs. They split the sentence containing the candidate pair into five segments: the sequence of tokens before/between/after the candidate pair, the trigger word, and the named entity of the candidate pair. Separate convolutional units were used to learn the latent representations of the five segments, and a final output layer was used to determine if the candidate pair is related or not.

MelaxTech continued the use of the BioBERT model re-trained on the patent texts released during the test phase. Similar to LasigBioTM, the input to their model is the sentence containing the candidate pair but only the candidate named entity is generalized by its semantic type in the sentences. Furthermore, rules were also applied in the post-processing step to recover false negative relations with a long distance, including relations across clauses and across sentences.

7.5 Summary of observations

The various approaches adopted by teams and the resulting performances have provided us with valuable experiences in how to address the tasks and what choices of methods are more suitable for our tasks.

Tokenization. In general, domain-specific tokenization tools perform better than tokenization methods that are for general English corpus. This is as expected since the vocabulary of chemical patents contains a large number of domain-specific terminology, and a machine can better understand and learn the characteristics of input texts if the texts are split into meaningful tokens. Another observation is that subword-based tokenization may contribute to overall accuracy. Chemical names are usually long, which make subword-based tokenization a suitable method for breaking down long chemical names. But further investigation is needed to support this claim.

Representation. Pre-trained embeddings are shown to be effective in enhancing system performances. Specifically, the Melaxtech and Lasige_BioTM systems are based on BioBERT model [23] and ranked the first and third place

in Task 1. The VinAI system leveraged embeddings pre-trained on chemical patents [48] and ranked second place. Character-level embeddings are also beneficial, shown by the ablation study in [11] and [27].

Model Architecture. The most popular choice of model is BERT [12], which is based on Transformer [43]. The model has demonstrated its effectiveness in sequence learning again. The Melaxtech system adopted this architecture and ranked first place in all three tasks. However, it is also worthwhile to note that the architecture of BiLSTM-CNN-CRF is still very competitive with BERT. The VinAI system ranked the first place in F_1-score when relaxed-match is used.

8 Conclusions

This paper presents a general overview of the activities and outcomes of the ChEMU 2020 evaluation lab. The ChEMU lab targets two important information extraction tasks applied to chemical patents: (1) named entity recognition, which aims to identify chemical compounds and their specific roles in chemical reactions; and (2) event extraction, which aims to identify the single event steps that form a chemical reaction.

We received registrations from 39 teams and 46 runs from 11 teams across all tasks and tracks, and 8 teams have contributed detailed system descriptions for their methods. The evaluation results show that many effective solutions have been proposed, with systems achieving excellent performance on each task, up to nearly 0.98 macro-averaged F_1-score on the NER task (and up to 0.99 F_1-score on a relaxed match), 0.95 F_1-score on the isolated relation extraction task, and around 0.92 F_1-score for the end-to-end systems. These results strongly outperformed baselines.

Acknowledgements

We are grateful for the detailed excerption and annotation work of the domain experts that support Reaxys, and the support of Ivan Krstic, Director of Chemistry Solutions at Elsevier. Funding for the ChEMU project is provided by an Australian Research Council Linkage Project, project number LP160101469, and Elsevier.

References

1. BRATEval evaluation tool. https://bitbucket.org/nicta_biomed/brateval/src/master/
2. CRF++ Toolkit. https://taku910.github.io/crfpp/, accessed: 2020-06-23
3. International Patent Classification. https://www.wipo.int/classifications/ipc/en/
4. Spacy tokenizer. https://spacy.io/api/tokenizer

5. Akhondi, S.A., Klenner, A.G., Tyrchan, C., Manchala, A.K., Boppana, K., Lowe, D., Zimmermann, M., Jagarlapudi, S.A., Sayle, R., Kors, J.A., et al.: Annotated chemical patent corpus: a gold standard for text mining. PLoS One **9**(9), e107477 (2014)
6. Akhondi, S.A., Rey, H., Schwörer, M., Maier, M., Toomey, J., Nau, H., Ilchmann, G., Sheehan, M., Irmer, M., Bobach, C., et al.: Automatic identification of relevant chemical compounds from patents. Database **2019** (2019)
7. Bregonje, M.: Patents: A unique source for scientific technical information in chemistry related industry? World Patent Information **27**(4), 309–315 (2005)
8. Carletta, J.: Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics **22**(2), 249–254 (1996)
9. Copara, J., Knafou, J., Naderi, N., Moro, C., Ruch, P., Teodoro, D.: Contextualized french language models for biomedical named entity recognition. In: Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes. pp. 36–48. ATALA (2020)
10. Copara, J., Naderi, N., Knafou, J., Ruch, P., Teodoro, D.: Named entity recognition in chemical patents using ensemble of contextual language models. In: Working Notes of CLEF 2020—Conference and Labs of the Evaluation Forum (2020)
11. Dao, M.H., Nguyen, D.Q.: VinAI at ChEMU 2020: An accurate system for named entity recognition in chemical reactions from patents. In: Working Notes of CLEF 2020—Conference and Labs of the Evaluation Forum (2020)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. pp. 4171–4186 (2019)
13. Hawizy, L., Jessop, D.M., Adams, N., Murray-Rust, P.: ChemicalTagger: A tool for semantic text-mining in chemistry. Journal of cheminformatics **3**(1), 17 (2011)
14. He, J., Nguyen, D.Q., Akhondi, S.A., Druckenbrodt, C., Thorne, C., Hoessel, R., Afzal, Z., Zhai, Z., Fang, B., Yoshikawa, H., Albahem, A., Cavedon, L., Cohn, T., Baldwin, T., Verspoor, K.: Overview of chemu 2020: Named entity recognition and event extraction of chemical reactions from patents. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020), vol. 12260. Lecture Notes in Computer Science (2020)
15. Jessop, D.M., Adams, S.E., Willighagen, E.L., Hawizy, L., Murray-Rust, P.: OSCAR4: a flexible architecture for chemical text-mining. Journal of cheminformatics **3**(1), 1–12 (2011)
16. Jurafsky, D., Martin, J.H.: Speech & Language Processing, 3rd edition, chap. Semantic Role Labeling and Argument Structure. Pearson Education India (2009)
17. Kim, J.D., Ohta, T., Pyysalo, S., Kano, Y., Tsujii, J.: Overview of BioNLP'09 shared task on event extraction. In: Proceedings of the BioNLP 2009 workshop companion volume for shared task. pp. 1–9 (2009)
18. Köksal, A., Hilal, D., Özkırımlı Elif, Özgür Arzucan: BOUN-REX at CLEF-2020 ChEMU Task 2: Evaluating Pretrained Transformers for Event Extraction. In: Working Notes of CLEF 2020—Conference and Labs of the Evaluation Forum (2020)
19. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the 18th International Conference on Machine Learning. pp. 282–289 (2001)

20. Lawson, A.J., Roller, S., Grotz, H., Wisniewski, J.L., Goebels, L.: Method and software for extracting chemical data. German patent no. DE102005020083A1 (2011)
21. Leaman, R., Gonzalez, G.: BANNER: an executable survey of advances in biomedical named entity recognition. In: Pacific Symposium on Biocomputing 2008, pp. 652–663. World Scientific (2008)
22. Lecun, Y.: Generalization and network design strategies. Technical Report CRG-TR-89-4, University of Toronto (June 1989)
23. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
24. Lowe, D., Mayfield, J.: Extraction of reactions from patents using grammars. In: Working Notes of CLEF 2020—Conference and Labs of the Evaluation Forum (2020)
25. Lowe, D.M., Sayle, R.A.: LeadMine: a grammar and dictionary driven approach to entity recognition. Journal of cheminformatics **7**(1), 1–9 (2015)
26. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. pp. 1064–1074 (2016)
27. Mahendran, D., Gurdin, G., Lewinski, N., Tang, C., T., M.B.: NLPatVCU CLEF 2020 ChEMU Shared Task System Description. In: Working Notes of CLEF 2020—Conference and Labs of the Evaluation Forum (2020)
28. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
29. Muresan, S., Petrov, P., Southan, C., Kjellberg, M.J., Kogej, T., Tyrchan, C., Varkonyi, P., Xie, P.H.: Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data. Drug Discovery Today **16**(23-24), 1019–1030 (2011)
30. Nakov, P., Hoogeveen, D., Màrquez, L., Moschitti, A., Mubarak, H., Baldwin, T., Verspoor, K.: Semeval-2017 task 3: Community question answering. arXiv preprint arXiv:1912.00730 (2019)
31. Ngai, G., Florian, R.: Transformation-based learning in the fast lane. In: Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies. pp. 1–8 (2001)
32. Nguyen, D.Q., Zhai, Z., Yoshikawa, H., Fang, B., Druckenbrodt, C., Thorne, C., Hoessel, R., Akhondi, S.A., Cohn, T., Baldwin, T., Verspoor, K.: ChEMU: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents. In: Proceedings of the 42nd European Conference on Information Retrieval. pp. 572–579 (2020)
33. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: An annotated corpus of semantic roles. Computational Linguistics **31**(1), 71–106 (2005)
34. Pattabhi, M.C., Rao., R., Lalitha Devi, S.: CLRG ChemNER: A Chemical Named Entity Recognizer @ ChEMU CLEF 2020. In: Working Notes of CLEF 2020—Conference and Labs of the Evaluation Forum (2020)
35. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 2227–2237 (2018)
36. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)

37. Ruas, P., Lamurias, A., Couto, F.M.: LasigeBioTM team at CLEF2020 ChEMU evaluation lab: Named Entity Recognition and Event extraction from chemical reactions described in patents using BioBERT NER and RE. In: Working Notes of CLEF 2020—Conference and Labs of the Evaluation Forum (2020)
38. Sætre, R., Yoshida, K., Yakushiji, A., Miyao, Y., Matsubayashi, Y., Ohta, T.: AKANE system: protein-protein interaction pairs in BioCreAtIvE2 challenge, PPI-IPS subtask. In: Proceedings of the second BioCreative challenge workshop. vol. 209, p. 212. Madrid (2007)
39. Schuster, M., Nakajima, K.: Japanese and korean voice search. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5149–5152. IEEE (2012)
40. Senger, S., Bartek, L., Papadatos, G., Gaulton, A.: Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents. Journal of Cheminformatics **7**(1), 1–12 (2015)
41. Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., Xu, H.: CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association **25**(3), 331–336 (2018)
42. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: BRAT: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107 (2012)
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
44. Verspoor, C.M.: Contextually-dependent lexical semantics (1997)
45. Verspoor, K., Nguyen, D.Q., Akhondi, S.A., Druckenbrodt, C., Thorne, C., Hoessel, R., He, J., Zhai, Z..: ChEMU dataset for information extraction from chemical patents. https://doi.org/10.17632/wy6745bjfj.1
46. Wang, J., Ren, Y., Zhang, Z., Zhang, Y.: Melaxtech: A report for CLEF 2020 – ChEMU Task of Chemical Reaction Extraction from Patent. In: Working Notes of CLEF 2020—Conference and Labs of the Evaluation Forum (2020)
47. Yoshikawa, H., Nguyen, D.Q., Zhai, Z., Druckenbrodt, C., Thorne, C., Akhondi, S.A., Baldwin, T., Verspoor, K.: Detecting Chemical Reactions in Patents. In: Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association. pp. 100–110 (2019)
48. Zhai, Z., Nguyen, D.Q., Akhondi, S., Thorne, C., Druckenbrodt, C., Cohn, T., Gregory, M., Verspoor, K.: Improving Chemical Named Entity Recognition in Patents with Contextualized Word Embeddings. In: Proceedings of the 18th BioNLP Workshop and Shared Task. pp. 328–338. Association for Computational Linguistics (2019)